*Environmental Health Perspectives*
*Vol. 96, pp. 67-75, 1991*

# Structure–Activity Relations: Maximizing the Usefulness of Mutagenicity and Carcinogenicity Databases

## by G. Klopman* and H. Rosenkranz†

**The most important criteria for the development and analysis of databases for elucidating the structural bases of toxicological activity include the integrity of the databases with respect to uniformity of the experimental protocol and interpretation of the test results and inclusion of chemicals representing different chemical classes and differing mechanisms of action. Within these criteria, it is demonstrated that when the chemicals are chosen at random, the larger the database, the better the predictivity of chemicals not included in the learning set. It is shown however, that when chemicals are selected on the basis of structural features, that a learning set of approximately 180 chemicals is as informative as a database consisting of 800 chemicals chosen at random.**

## Introduction

Whereas most current studies, including those reported in this symposium, deal with the classification of information, our approach is to rationalize the available data and thus permit extrapolation to molecules that have as yet not been tested. The eventual outcome of such an approach is to help optimize ongoing studies such that they, in turn, will provide a maximal amount of information of a mechanistic and predictive nature.

A characteristic of the available databases of direct concern to us is that they are composed of a wide variety of chemicals with diverse structures, e.g., polycyclic aromatic hydrocarbons, nitrosamines, halogenated hydrocarbons, and dyes. They are thus not amenable to traditional quantitative structure–activity relationship (QSAR)-type studies, which require congeneric databases. To overcome this obstacle, we have been investigating how knowledge-based systems could be useful.

In this survey we discuss our experience with using available databases of carcinogenic, mutagenic, and related biological end points to establish such a resource undertaken for mechanistic as well as predictive purposes. We stress that our analysis is, of course, influenced by the structure–activity relationship (SAR) methodology we employ, namely, CASE (*1,2*). However, CASE, as an artificial intelligence/expert system, is probably one of the most advanced SAR methods available and a harbinger of the developments that are to come to this methodology in general. In this connection, it is germane to list some of the unique features that were incorporated into this program:

1. In order to make the system truly effective, it had to be independent of operator-formulated questions. These by definition, are finite and simplistic and usually based upon previous knowledge (i.e., they may be biased). Thus, we required the system to generate all of the possible structural "descriptors" automatically without operator input.
2. Because we felt that the biological activity is dependent on molecular subunits usually larger and more complex than those considered by chemists as simple functionalities, we required that our system be capable of handling and identifying relatively large molecular moieties.
3. We realized that most SAR systems are based on the analysis of congeneric databases. e.g., nitroarenes, aromatic amines, halogenated hydrocarbons, etc. This, however, involves the arbitrary assignment of chemicals to certain classes and, when dealing with databases containing various chemical classes, might *a*) introduce bias in the selection process and *b*) make the emerging classes too small for adequate structural analysis. Moreover, if biological activity is derived inherently from structural features, by "artificially" separating chemicals into restricted classes, we might lose informational content. Accordingly, we needed a system that was able to handle, in a single database, molecules of very different chemical types.
4. We also wanted to include in our system the capability of updating the "descriptors" as information on new molecules became available, i.e., the system had to be self-learning.
5. We required the resulting analyses not only to be predictive but also to provide clues as to possible mechanisms of biological activity.
6. Finally, we wanted the system to handle biological end points that may result from a single mechanism, a com-

*Department of Chemistry, Case Western Reserve University, Cleveland, OH 44106.

†Department of Environmental and Occupational Health, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA 15261.

Address reprint requests to G. Klopman, Department of Chemistry, Case Western Reserve University, Cleveland, OH 44106.

bination of concerted or sequential actions, or two or more independent mechanisms.

The program we constructed to meet these requirements is called CASE and, in its latest and greatly improved version, MULTICASE.

## Nature of the Database

What are some of our conclusions relating to the requirements of the composition of a database to be useful for SAR studies? Because of the growing sophistication of computer-based SAR systems and the investment of time as well as CPU cost, it is important to begin with carefully evaluated databases. Some of the requirements for such databases are self-evident: The database must be obtained using standard protocols for which the quality of the data must be monitored. This is borne out by experience with some of the cytogenetic end points that are contained in the Gene-Tox Program database as compared to those in the National Toxicology Program (NTP) compilations. Thus, even using the peer-review process inherent in the Gene-Tox Program, the SAR methodology could not be applied with great success to some of the Gene-Tox databases (unpublished results), whereas the NTP databases allowed thorough analyses of the structural features of the cytogenetic activities (*3*). Our analysis of the data led us to conclude that the difference came from the quality control, which was assured in the NTP protocol but could not be controlled in the Gene-Tox Program. This is due to the fact that the latter relied on published data, and, moreover, the chemical purity (or rather impurity) was not generally known. Nevertheless, CASE can handle some "fuzziness" in the data because it is based on the statistical evaluation of the importance of substructure rather than on a quantitative relation.

The second, also perhaps self-evident, point relates to the interpretation of the data. Both the biological and statistical standards used for interpreting the test results must be adhered to rigidly and must, of course, be spelled out initially. This is especially important in the manner in which results are expressed when a continuous activity scale is used, e.g., revertants per nanomole or milligrams per kilogram per day. Although there are computer routines to scale and model cutoffs, this judgment should not be delegated to a computer. In fact, we find that the human expert is essential to determine the boundary between inactive and marginally active chemicals and between marginally active and active chemicals. The scaling can then be done by the computer following these initial boundary settings.

In addition to defining the accepted definitions of the end points with respect to databases, the purpose of the analysis must also be defined *a priori*. Thus, for different mechanistic purposes we might use different mutagenicity databases or different collections of data derived from these compilations. For example, *a*) mutagenicity in a specific Salmonella tester strain in the absence of S9 might be used. The purpose of such an analysis would be perhaps to study the structural basis of the mutagenicity of nitrated polycyclic aromatic hydrocarbons, which are maximally expressed in strain TA98 in the absence of exogenous metabolic activation (*1,4*). *b*) The activity might be in Salmonella TA100 or in TA98 in the presence of S9. Thus, for the aromatic amines, this would allow the determination of specific combinations that would be informative. In fact, using a database of only TA100 in the presence of S9 as against one of only TA98 in the presence of S9, the effect of mutagenic specificity by aromatic amines at a single guanine-cytosine (G–C) base pair (as in TA100) as opposed to the specificity at a series of alternating G–C pairs (as in TA98) was amenable to analysis (*5,6*). *c*) On the other hand, if we wish to study and evaluate how Salmonella mutagenicity compares in its predictivity to the results of rodent carcinogenicity bioassays, then for the Salmonella data we might define as a positive response a response in any one of the tester strains obtained either in the presence or absence of S9 (*2,7*). Under these conditions, of course, it must be ascertained that before a chemical is designated as negative that it has indeed been tested in the complete panel of tester strains both in the presence and absence of S9. Obviously, this approach, which has also been used to assess the predictivity of the Salmonella assay for carcinogenicity in rodents (*8–10*), assumes that the structural determinants responsible for activity in different strains are identical and equally related to carcinogenicity. In fact, we know that these are oversimplifications.

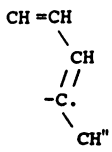
## Nature of the Chemicals in the Database

The nature of the chemicals represented in the database is probably the most difficult to address. In view of the fact that CASE can handle noncongeneric databases, then some of the requirements are easy to state but difficult to implement: A database (i.e., learning set) should contain representatives of various chemical classes and chemicals that cover the spectrum of mechanisms that induce a particular biological end point (*11*). To satisfy such a requirement, it might seem a truism to say that a variety of chemical classes need to be represented in the data base. However, this is more easily said than done. Looking at the problem from a chemical point of view, using, for example, the experience of the Gene-Tox Program, initially, in excess of 60 chemical classes were defined (*12*). This resulted in a situation that for SAR studies, there were very few representatives per chemical class such that SAR analyses were not feasible. Subsequently, the number of chemical classes was reduced to 30 (*13*). Still, this left many chemical classes underrepresented for the purpose of SAR analysis. Moreover, such separations imply that we know those structural features that are necessary for biological activity. This, of course, involves a selection bias. On the other hand, a system like CASE selects its own descriptors, and these in turn allow it to bypass the traditional chemical classes (Table 1).

It is of interest that by and large the CASE program usually identifies not more than approximately 12 to 15 significant biophores in a noncongeneric database (Table 2). This means that there are usually a sufficient number of representative chemicals in each of the biophore classes that are selected. Additionally, CASE also identifies biophobes, i.e., functionalities that contribute to a lack of activity. These might be considered as "non-alerts" in the scheme of Ashby (*14*).

In creating databases, if one has unlimited funds, one might decide to test as many chemicals as possible and enter the results into the database. Under such conditions, one need not be concerned about redundancy of chemical structures. Thus, one would not be worried that there might be too many nitrofurans in a database, as the logic of the CASE program sees to it that the biophore associated with nitrofurans will achieve significance when a certain preset probability value has been reached (e.g.,

**Table 1. Molecules sharing a common biophore.[a]**



*o*-Phenanthroline
D & C Yellow 11
9-Nitroanthracene
1-Nitro-2-methylnaphthalene
C.I. Pigment Red 3
C.I. Pigment Red 23
D & C Red 9
8-Hydroxyquinoline
C.I. Solvent Yellow 14
*N*-Phenyl-2-naphthylamine
Benzo(*a*)pyrene
3-Methylcholanthrene
Benz(*a*)anthracene
7,12-Dimethylbenz(*a*)anthracene
Phenanthrene
Benzo(*f*)quinoline (beta-Na)
1-Nitronaphthalene
Naphthalene
Quinoline
2-Naphthylamine
Coumarin
Anthracene
Pyrene
1-Naphthylamine
Benzo(*e*)pyrene
1,8,9-Trihydroxyanthracene
1-Naphthylisothiocyanate
2-Anthramine
7,9-Dimethylbenz(*c*)acridine
*N*-(1-Naphthyl)ethylenediam
7-Bromomethyl-12-methylbenz(*a*)anthracene

[a]Note that polycyclic aromatic hydrocarbons, aromatic amines, nitroarenes, and others contain this biophore, which has been identified as significant with respect to mutagenicity in Salmonella ($p = 0.0001$).

**Table 2. Some of the biophores and biophobes contributing to the probability of carcinogenicity in rodents.[a]**

| Fragment | Number of fragments | Inactive | Marginally active | Active | Probability |
|---|---|---|---|---|---|
| Activating[b] | | | | | |
| $NH_2$—C= | 53 | 12 | 5 | 36 | 0.000 |
| Cl—CH2— | 19 | 5 | 2 | 12 | 0.058 |
| C''—O—C= | 5 | 0 | 0 | 5 | 0.031 |
| O—$CH_2$—CH— | 6 | 0 | 1 | 5 | 0.031 |
| C=CH—C=C— | 49 | 11 | 6 | 32 | 0.001 |
| CH=C—C=CH— | 54 | 14 | 7 | 33 | 0.005 |
| CH=C•—CO—C•= | 5 | 0 | 0 | 5 | 0.031 |
| $CH_2$—$CH_2$—$CH_2$—CH— | 5 | 0 | 0 | 5 | 0.031 |
| O—CO—C=CH— | 9 | 0 | 3 | 6 | 0.020 |
| Inactivating | | | | | |
| OH—CH— | 9 | 8 | 1 | 0 | 0.004 |
| S—CH— | 5 | 4 | 1 | 0 | 0.063 |

[a]This table lists the distribution of the major fragments among active, inactive, and marginally active molecules. This distribution is used to predict the likelihood that the presence (or absence) of the fragment contributes to carcinogenicity. Also listed are the probability values associated with the fragments.
[b]C• indicates a carbon atom common to two rings. C'' indicates the carbon is attached by a double bond to an outside substituent.

$p < 0.05$), at which point that biophore will be flagged and identified. Having more representative molecules containing that biophore in the pool will not contribute overwhelmingly to the predictivity. Similarly, using the same reasoning, if the database contains too great a prevalence of active molecules, this will not affect the identification and performance of the predictive biophores that are identified provided that there are a sufficient number of molecules in the database.

## Congeneric versus Noncongeneric Databases

Heretofore the majority of SAR methods were designed for the study and prediction of congeneric databases such as those containing polycyclic aromatic hydrocarbons, nitroarenes, aromatic amines, etc. One of the breakthroughs provided by CASE is its ability to analyze none congeneric databases (i.e., mixed databases). The question then facing us is: Everything else being equal, how do the two types of databases compare with respect to predictivity?

Let us take the NTP Carcinogen Database as an example. In one of our analyses we had approximately 250 chemicals, of which 53 were aromatic amines (*15,16*) (Table 3). As an exercise, we then selected the 53 aromatic amines and used them to construct a congeneric database. We then used the two databases, the noncongeneric and the congeneric one to study the predictivity of each of them for aromatic amines. The results indicate that both databases were highly predictive of the carcinogenicity of aromatic amines (Table 4). However, further analysis indicated that the noncongeneric database was significantly more predictive than the congeneric one. This appears to be derived from the fact that the biophores selected by CASE may cut across chemical species and could, for example, have been derived in the case of the noncongeneric database not only from aromatic amines but possibly from related nitroarenes. In the case of the congeneric database consisting only of aromatic amines, such biophores would not necessarily be identified (Table 5). Thus, the noncongeneric database may indeed be superior even when a sufficient number of congeneric database chemicals are present therein because CASE can learn from related molecules which, for example, may yield the same metabolic intermediates (e.g., *N*-arylhydroxylamines), which are derived from both arylamines and nitroarenes by oxidative and reductive pathways, respectively.

**Table 3. Relationship between structure of aromatic amines and carcinogenicity in rodents.**

| Fragment | Total | Inactive | Equivocal | Active | % Actives and equivocals |
|---|---|---|---|---|---|
| All molecules | 252 | 88 | 29 | 135 | 65.1 |
| $NH_2$—C= | 53 | 12 | 5 | 36 | 77.4 |
| $NH_2$—C=CH—CH=[a] | 38 | 10 | 5 | 24 | 76.3 |
| $NH_2$—C=C—CH=CH—[b] | 19 | 3 | 0 | 16 | 84.2 |
| $NH_2$—C=CH—CH= C—C—$NH_2$[b] | 4 | 0 | 0 | 4 | 100 |

[a]This fragment is not associated with an increased probability of carcinogenicity.
[b]CASE identified these aromatic amine-derived fragments as associated with an increased probability of carcinogenicity.

**Table 4. Comparison of congeneric versus noncongeneric database in the prediction of the carcinogenicity of aromatic amines.**

| | Noncongeneric[a] | Aromatic amines |
|---|---|---|
| Number | 252 | 53 |
| Carcinogens | 65.1% | 77.4% |
| Correct predictions | 96.2% | 88.7% |
| Expected (random) predictions | 56.4% | 56.4% |
| $\chi^2$ | 44.22 | 29.04 |

[a]Contains the 53 aromatic amines.

**Table 5. Some nonaromatic amines that share biophores CH=C(R)—C(R)=CH— with aromatic amines.[a]**

| Chemical | Carcinogenicity[b] | Mutagenicity (Salmonella) |
|---|---|---|
| C.I. Disperse Yellow 3 | A | + |
| HC Blue 1 | A | + |
| Nitrofen | A | + |
| D & C Red 9 | B | + |
| 3,3'-Dimethoxybenzidine-4,4'-diisocyanate | B | + |
| 2,4-Dinitrotoluene | B | + |
| 3-Nitro-*p*-acetophenetide | D | + |
| HC Blue 2 | NC | + |
| Tetrachlorvinphos | A | − |

[a]R is not a hydrogen. This biophore was identified in the total database consisting of 253 chemicals. It was not found to have significance in the database consisting only of aromatic amines.

[b]Classifications of Ashby and Tennant (*9*). A, induces tumors in rats and mice; B, carcinogenic to only one species but induces cancers at two or more sites; D, carcinogenic at a single site in a single species; NC, noncarcinogenic.

## How Many Chemicals Are Needed in a Database?

Obviously, we do not possess unlimited resources, and the number of chemicals that can be tested is by necessity limited. Therefore, we might ask the question, how many chemicals are needed in the learning set when it consists of noncongeneric molecules? This is a question that needs to be decided when testing programs are set up and the data used for SAR analyses.

In Table 6 we show the predictivity of using different numbers of chemicals in the learning set. The chemicals in the learning set were selected at random from the NTP Salmonella Mutagenicity Database. The biophores identified using the different databases were then used to predict the mutagenicity of a panel of 100 chemicals, not in the learning set, but for which the test results were available ("diagnostic tester set").

The results of the analysis clearly show that the predictive performance of CASE improves with an increase in the number of chemicals in the learning set (Table 6). However, generating databases consisting of such large numbers of chemicals is costly, even if we restrict the testing to the Salmonella mutagenicity assay only. Obviously, if we use experimental systems that are more labor intensive or which involve large numbers of animals, the cost will rapidly become prohibitive. Thus, it is not surprising that very few rodent cancer bioassays have been repeated given a cost which may exceed $1 million per assay.

In search for a method to decrease the number of chemicals that require testing (i.e., to limit the size of the database), we explored a number of possibilities. We devised an efficient procedure, which is dependent on another feature of CASE. Indeed, CASE predictions can take a number of forms: *a*) a chemical can be predicted to be active because it contains a biophore (Fig. 1); *b*) a chemical can be predicted to be inactive because it contains a biophobe (Fig. 2); *c*) a molecule can be predicted to be inactive because it lacks a recognizable biophore and/or biophobe. That later prediction is due to the fact that fragmentation of the molecule yields fragments that had been seen before but had been judged to be trivial with respect to biological activity (Fig. 3); *d*) an additional possibility is that the CASE program has identified an unrecognized functionality (i.e., unknown). This refers to the presence of a fragment that has not been documented among the collection of fragments generated (Figs. 4–6). The presence of such a fragment introduces a note of inconclusiveness into the prediction, which appears to be the major reason for decreased predictive performance (*17*). This uncertainty might, in fact, be a direct function of the number of chemicals in the learning set. That is, the more chemicals there are in the learning set, the less the chance that this message will appear (Table 7).

**Table 6. Predictive performance of databases containing increasing numbers of randomly selected chemicals.[a]**

| *n* | Concordance, % | $\chi^2$ |
|---|---|---|
| 25 | 60.2 | 5.89 |
| 50 | 58.9 | 3.78 |
| 100 | 68.2 | 14.67 |
| 243 | 75.3 | 27.4 |
| 820 | 81.0 | 31.3 |

[a]All of the subsets of chemicals are contained within the set of 820. The chemicals were chosen at random from among the 820 molecules in the database except for the 243 chemicals, also a subset of the 820 that represents chemicals on which rodent cancer bioassays were performed (*9,10*).

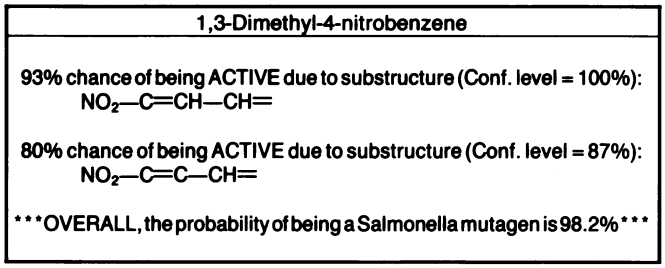



FIGURE 1. CASE prediction of the mutagenicity in Salmonella of 1,3-dimethyl-4-nitrobenzene.

## Can SAR Concepts Reduce the Number of Chemicals That Need to Be Tested to Generate a Useful Database?

As shown above, the larger the database, the better the predictive performance. However, the question then still is are we really interested in generating, at a great cost, databases that allow 85, 90, 99, or even 99.9% concordances between experimental results and predictions, especially if the actual experimental data are only 85% reproducible, as appears to be the situation with the NTP Salmonella Mutagenicity Database? In fact, a number of analyses have indicated that the limit of the predictivity of the CASE program for mutagenicity in Salmonella is

approximately 80% (*17*). Moreover, the reproducibility of the rodent carcinogenicity bioassay is largely unknown since so few of the chemicals have been tested more than once. We ought to aim for economy and reliability.

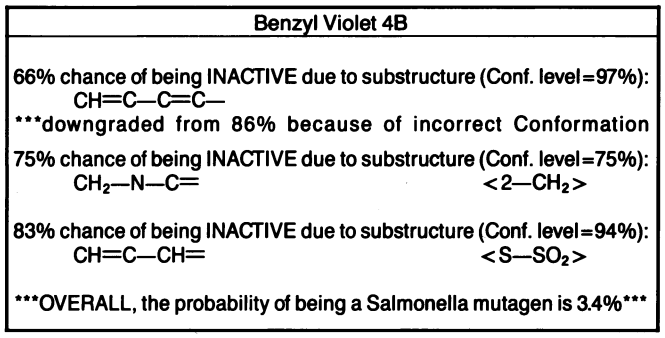



FIGURE 2. CASE prediction of the lack of mutagenicity of benzyl violet 4B.

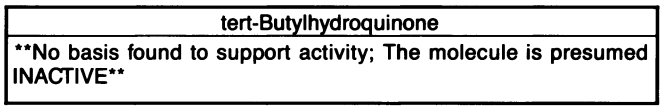



FIGURE 3. CASE prediction of the lack of mutagenicity of *tert*-butylhydroquinone. Fragmentation of this molecule does not yield fragments that have been identified as biophores of biophobes.

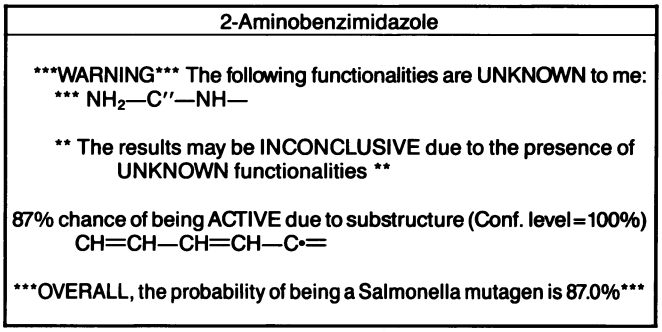



FIGURE 4. CASE prediction of the mutagenicity of 2-aminobenzimidazole. CASE recognized a biophore that leads to an 87% probability of mutagenicity. However, this conclusion must be moderated by the fact that this molecule contains a fragment that has not been seen before and that could be a biophobe.

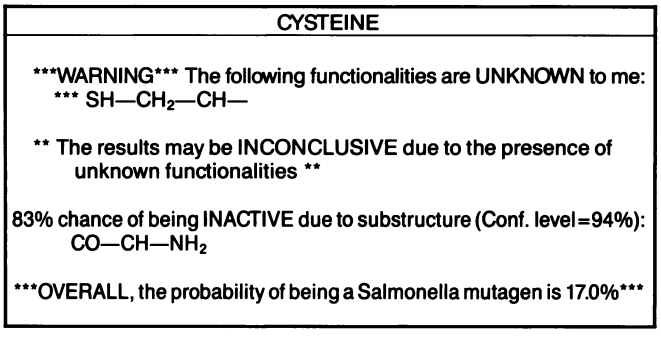



FIGURE 5. CASE prediction of the lack of mutagenicity of cysteine due to the presence of a biophobe. However, this conclusion is moderated by the fact that cysteine contains a fragment not seen before, which might be a biophore.

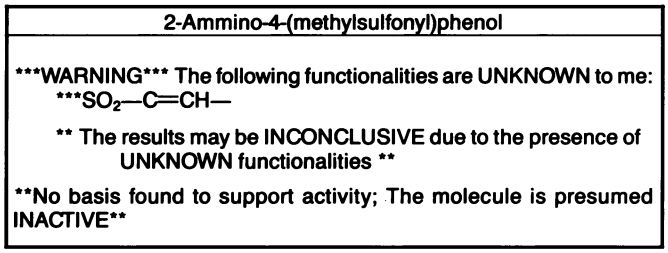



FIGURE 6. A CASE prediction of lack of mutagenicity due to the fact that fragmentation of the molecule does not lead to the generation of either a biophore or a biophobe. However, this prediction must be moderated by the fact that a fragment, heretofore not seen, has been recognized, and it could be a biophore.

**Table 7. Effect of size of the learning set on the number of inconclusive predictions.[a]**

| Number of chemicals in learning set | Inconclusive predictions, %[b] |
|---|---|
| 25 | 61 |
| 50 | 54 |
| 100 | 47 |
| 250 | 33 |
| 820 | 22 |

[a]The diagnostic tester set consisted of 100 molecules that were not present in the learning sets.

[b]Number of predictions in the diagnostic tester that contain fragments that had not been seen before (see Figs. 4–6) and which therefore may be inconclusive.

Our analysis of the performance of CASE suggests that perhaps we could use the uncertainty factor (see above) in the design of a database that will contain the fewest chemicals that need to be tested. To test this hypothesis, the following protocol was devised and tested: *a*) a list of chemicals including different chemical classes, different uses, and different levels of production (*18*) is selected; *b*) from among this list, we might select, at random, say 100 chemicals that will be tested (or which already may have been tested). This forms the original learning set (set 1); *c*) this original learning set is analyzed by CASE and the biophores and biophobes are identified; *d*) another set of 50 randomly selected chemicals (excluding those chemicals in set 1), as yet untested, are run against this original learning set; *e*) those chemicals which, in step *d*, yield a prediction that includes the uncertainy message are then selected for testing as mutagens in Salmonella. The results of the tests are then included in a new learning set (set 2, which includes the chemicals in set 1); *f*) this procedure is performed iteratively.

To evaluate the effectiveness of the selection procedure at each step, the predictivity of the database is evaluated by challenging it with the hundred chemicals not included in the learning sets but for which test results are available ("diagnostic tester set"). The results of these analyses clearly indicate (Table 8) that by careful selection of chemicals for testing, the number of chemicals that need to be tested to generate a learning set adequate for SAR predictions can be greatly reduced. This is accompanied by a corresponding decrease in cost. In fact, a comparison of Tables 6 and 8 clearly indicates that a database of approximately 180 carefully selected chemicals is as predictive as a database consisting of in excess of 800 chemicals that have not been selected by the criteria. Similar results were obtained (*19*) using the Gene-

**Table 8. Predictive performance of database containing increasing numbers of selected chemicals.**

| Set | *n* | Concordance, % | $\chi^2$ |
|---|---|---|---|
| 1 | 100[a] | 69.3 | 6.36 |
| 2 | 119 | 77.0 | 19.75 |
| 3 | 142 | 77.6 | 21.89 |
| 4 | 160 | 74.0 | 16.59 |
| 5 | 179 | 80.8 | 28.11 |
| 6 | 198 | 78.0 | 22.76 |
| 7 | 210 | 80.6 | 26.77 |
| 8 | 219 | 78.1 | 22.59 |

[a]The first 100 chemicals are presumed to represent a random assortment of molecules. They were not selected by the procedure described here. Each subsequent set contains an increment of molecules selected from sets of 50. Thus, set 2 contains the previous 100 molecules (set 1) plus 19 selected from among a set of 50. Similarly, set 3 contains the previous 119 moleclues (set 2) plus 23 molecules selected from among another set of 50. This procedure is used iteratively. The selection of molecules is described in the text.

Tox Salmonella database, which not only contains a different collection of chemicals but also a higher prevalence of mutagens (78.5% versus 36.5%).

## Quantitative versus Binary Databases

There are two major ways of expressing the results of SAR analyses: *a*) as active, marginally active, and inactive. This, *a priori*, involves a prejudgment involving the human expert who then sets the boundaries as to what is considered a positive, marginal, and negative result. Summarizing the results in this manner and applying the CASE program leads to the generation of fragments that are involved in the probability of a certain biological activity (e.g., mutagenicity, carcinogenicity) and therefore provides a possible procedure for risk identification for testing prioritization. Indeed, such probabilities appear to be related to the biological properties. Thus, the degree of the carcinogenicity of a chemical as defined by Ashby and Tennant (*9*) appears to be related to the probability of carcinogenicity: an overall high probability is associated with chemicals that cause cancer in both rats and mice at multiple sites of both sexes (Table 9). *b*) When, however, results are expressed in a continuous scale (e.g., $TD_{50}$ in milligrams per kilogram per day or mutagenicity in revertants per nanomole), this permits two independent analyses to be carried out: a probability of carcinogenicity, identical to the procedure described above and a QSAR analysis, which leads to the identification of biophores and biophobes associated with potency (Table 10). The latter analysis leads to a second prediction, that of potency (Fig. 7).

Thus, from the fragments associated with the mutagenic potency of nitroarenes (Table 10), we can calculate the projected activity of a chemical where

$$\text{CASE activity} = 9.208 + nF + 1.22 \log P$$

where $n$ is the number of times a fragment occurs in the molecule, $F$ is the CASE activity associated with that fragment (Table 10), and $P$ is the $n$-octanol/water partition coefficient. Accordingly, for 1,6-dinitropyrene (Fig. 7), biophore B (which is present twice) is the same as biophore 7 (Table 10), which is associated with a mutagenic potency of 21.576 units, and, moreover, 1,6 dinitropyrene also contains two copies of biophore 3 (17,093 units) (Table 10).

**Table 9. Relationship between the probability of carcinogenicity and carcinogenic classification.**

| | Rodents | | Rats | | Mice | |
|---|---|---|---|---|---|---|
| Group[a] | *n* | Probability | *n* | Probability | *n* | Probability |
| A | 64 | 75.0 ± 14.0 | | | | |
| B | 19 | 73.9 ± 14.2 | 50 | 66.3 ± 11.0 | 46 | 67.2 ± 17.3 |
| C | 27 | 67.9 ± 16.9 | 21 | 59.6 ± 15.3 | 34 | 65.3 ± 10.9 |
| D | 25 | 70.8 ± 12.9 | 29 | 63.2 ± 17.7 | 19 | 62.3 ± 16.7 |
| E | 29 | 45.0 ± 30.9 | 18 | 7.4 ± 18.8 | 17 | 18.7 ± 27.1 |
| NC | 88 | 74.8 ± 13.1 | 134 | 9.0 ± 17.6 | 136 | 8.1 ± 14.8 |
| A + B | 83 | 74.8 ± 14.1 | 50 | 66.3 ± 11.0 | 46 | 67.2 ± 17.3 |
| C + D | 52 | 69.1 ± 11.6 | 50 | 61.7 ± 16.8 | 53 | 64.3 ± 13.3 |
| A11 | 252 | 51.2 ± 30.2 | 252 | 30.7 ± 31.7 | 252 | 34.1 ± 32.6 |

[a]The classification of Ashby and Tennant (*9*) was adopted. In that scheme, chemicals in group A induce tumors in rats and mice. Group B includes chemicals that are carcinogenic to only one species but which include cancers at two or more sites. Group C consists of chemicals that are carcinogenic at only a single site in both sexes of a single species. Group D contains chemicals carcinogenic at a single site in a single species. Group E is adequately studied chemicals for which only equivocal evidence of carcinogenicity was observed. NC, noncarcinogens.

**Table 10. Some biophobes associated with the mutagenic potency of nitroarenes.[a]**

| QSAR fragments | | Number of chemicals | QSAR | |
|---|---|---|---|---|
| $NO_2$—C=CH—C·= | | 45 | 11.382+++ | 1 |
| CH=C·—C=CH— | <3—NO> | 3 | 17.093+ | 2 |
| CH=C·—C=CH— | <3—$NO_2$> | 49 | 17.093+++ | 3 |
| OH—N—C=CH— | <3—CH=> | 1 | 5.390 | 4 |
| NO—C=CH—CH=C·— | | 3 | 21.576+ | 5 |
| $NO_2$—C=CH—CH=C— | | 15 | 8.754+++ | 6 |
| $NO_2$—C=CH—CH=C·— | | 57 | 21.576+++ | 7 |
| CH=C—C=CH—C= | <3—$NO_2$> | 10 | 3.982+++ | 8 |
| $NO_2$—C=C—CH=CH—C= | | 10 | 3.982+++ | 9 |
| CH=CH—C=CH—C=C— | <5—$NO_2$> | 10 | 3.982+++ | 10 |
| $NO_2$—C=CH—C=CH—CH=C— | | 10 | 3.982+++ | |

[a]Constant = 9.208.

**1,6-Dinitropyrene**

(A) 97% chance of being ACTIVE due to substructure (Conf. level = 100%): CH=CH—C·=C—CH=CH—C·=
(B) 95% chance of being ACTIVE due to substructure (Conf. level = 100%): $NO_2$—C=CH—CH=C·—
(C) 92% chance of being ACTIVE due to substructure (Conf. level = 100%): CH=C·—C=CH—

***OVERALL, the probability of being a Salmonella mutagen is 100%***
**The compound is predicted to be EXTREMELY active (92) **

FIGURE 7. Prediction of the mutagenicity and mutagenic potency of 1,6-dinitropyrene.

Another feature of CASE is the estimation of the log $P$, which for 1,6-dinitropyrene is 4.7968. Accordingly, the CASE activity of 1,6-dinitropyrene is

$$9.208 + [2(17.093)] + [2(21.576)] + 4.7968 = 92$$

CASE activity is expressed in a log scale; an activity of 92 is equivalent to 200,000 revertants/nmole. This, in fact, goes beyond mere risk identification, which is based on the probability of activity, because the prediction also involves a measure of potency that can be used in a quantitative risk assessment. Ob-

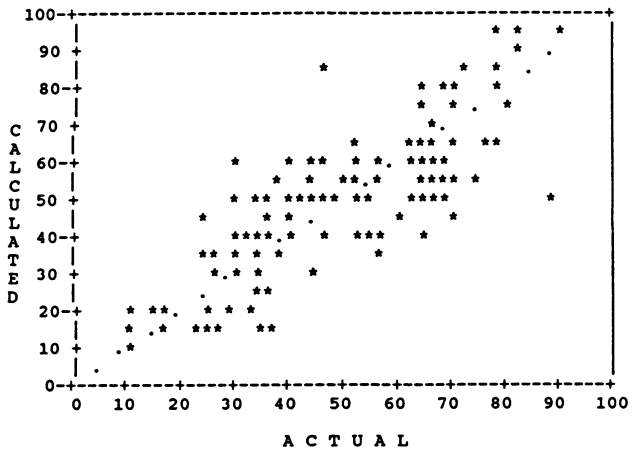



FIGURE 8. Relationship between experimental and predicted mutagenic activity for a series of nitroarenes. The activities are expressed in CASE activity units, which is a logarithmic scale; $r^2 = 0.8371$.

viously, before using such potency values, they must be compared against the experimentally obtained values (Fig. 8).

In addition to allowing an estimation of the potency of an unknown chemical, the identification of the biophores associated with potency permits an assessment of the relative role of different biophores in activity. Thus, as a result of the logarithmic nature of the activity scale, it can be calculated that biophore 7 (Table 10) contributes 88% of the mutagenicity of 1,6-dinitropyrene; i.e, it is the major contributor to potency. Such analyses have obvious mechanistic implications (*20*). Thus, studies seeking to investigate the basis of the potent mutagenicity of this chemical should concentrate on that portion of the molecule.

## Comparison of Some Carcinogenicity Databases

We have performed analyses of a number of different rodent carcinogenicity databases including the NTP rodent bioassay (*9,10*) and the compilation of Gold and associates (*21-23*) which includes $TD_{50}$ values. Analyzing these databases allowed us to identify the structural features responsible for the probability of carcinogenicity (i.e., the qualitative aspect of the analysis). Moreover, by analyzing carcinogenicity for only the mouse or only the rat, we were able to identify biophores characteristic for each of these activities as well as biophores common to both the rat and the mouse. These have mechanistic implications that will not be described here (see below).

Applying the QSAR CASE analysis to the database assembled by Gold et al. (*21-23*), which includes $TD_{50}$ values, indicated that the data could be used to generate QSAR relationships for individual databases (i.e., mouse and rat separately) which, in turn, can be used to project carcinogenic potencies based upon the QSAR contribution of individual biophores (Table 11). [The $TD_{50}$ value is defined as the lifetime dose (milligrams per kilogram per day) that reduces by one-half the lifetime chance of remaining tumor-free (*24*). Thus, carcinogenic potency can be projected in a manner similar to that described for the mutagenicity of 1,6-dinitropyrene (see above).

## Validation

Before using the biophores and biophobes for predictive and mechanistic studies, a number of controls need to be performed. Routinely, from the available database, a set of randomly selected chemicals is removed before the CASE analysis. These chemicals are then used as a tester set to test the predictivity of the data set. Subsequently, of course, these chemicals can be added back to the learning set and the CASE analysis performed again.

Additionally, when analyzing biological activities such as mutagenicity, cytogenotoxicity, and carcinogenicity, we also assembled a database of naturally occurring physiological chemicals (amino acids, sugars, lipids, purines, pyrimidines, vitamins, etc.). These chemicals are expected to be negative. However, on occasion, we have found that some databases led to predictions that a significant fraction of physiological chemicals induced some end points, e.g., sister chromatid exchange. Such findings cast doubt on the relevance of such assays as predictors of carcinogenicity (*25*).

## Data Management

Finally, a data management system must be in place to keep track of the various predictions that are made in the course of these analyses. Additionally, this will enable testing correlations of predictions. Thus, such a data management system permitted

**Table 11. Some biophores and biophobes contributing to the carcinogenic potency.[a]**

| QSAR fragments[b] | Number of fragments | Inactive | Marginally active | Active | QSAR |
|---|---|---|---|---|---|
| NH—CH— | 5 | 4 | 1 | 0 | −11.445− − |
| OH—CH | 9 | 8 | 1 | 0 | −11.445− − − |
| S—CH— | 5 | 4 | 1 | 0 | −11.445− − |
| Cl—CH= | 4 | 0 | 0 | 4 | 19.531+ + |
| Cl—$CH_2$— | 19 | 5 | 2 | 12 | 10.057+ + |
| C″—O—C= | 5 | 0 | 0 | 5 | 30.867+ + + |
| C=CH—C=C— | 49 | 11 | 6 | 32 | 6.387+ + + |
| CH=C·—CO—C·= | 5 | 0 | 0 | 5 | 8.502+ + + |
| NH2—C=C—CH=CH— | 19 | 3 | 0 | 16 | 22.936+ + + |
| Cl—C=C—CH=CH— | 6 | 5 | 0 | 1 | −13.869− − |
| CH=CH—C=CH—CH= <3—N> | 8 | 2 | 0 | 6 | 9.863+ + |
| CH=CH—CH=CH—C·=C—CH= | 4 | 4 | 0 | 0 | −19.938− − |

[a]Constant = 23.818. The QSAR activity is used to calculate potency using the equations: QSAR activity = $23.818 + n_f F_f n_b F_b$ where $n$ denotes the number biophore $F_f$ or biophobe $F_f$ that is present in the molecule.

[b]C· indicates a carbon atom common to two rings. C″ indicates the carbon is attached by a double bond to an outside substituent.

**Table 12. Expected distributions of test results among a population of random molecules.[a]**

| Assay system | Positive responses, % |
|---|---|
| Rodent carcinogens | 56 |
| Mutagenicity in Salmonella | 39 |
| Chromosomal aberrations (CHO cells) | 50 |
| Sister chromatid exchanges (CHO cells) | 71 |
| Rodent carcinogen and mutagenicity | 39 |
| Rodent carcinogenicity and sister chromatid exchanges | 44 |
| Rodent carcinogenicity and chromosomal aberrations | 33 |
| Sister chromatid exchanges and chromosomal aberrations | 44 |

[a]These distributions are based upon the analyses of approximately 1150 chemicals representing many sources and uses.

**Table 13. Effect of the prevalence of carcinogens among molecules on the results of short-term tests: predicted positive responses of short-term tests.**

| | Proportion of carcinogens among population of molecules | | | | | |
|---|---|---|---|---|---|---|
| Assay | 0% | 10% | 20% | 45% | 65% | 100% |
| Salmonella | 13.9% | 17.3% | 20.9% | 27.5% | 34.0% | 44.2% |
| Sister chromatid exchanges | 47.0% | 49.4% | 50.8% | 56.6% | 62.8% | 68.5% |
| Chromosomal aberrations | 23.3% | 25.7% | 27.9% | 35.3% | 41.6% | 50.5% |

us to determine how often carcinogens are predicted to be mutagens or how often Salmonella mutagens are predicted also to induce chromosomal aberrations (Table 12).

Such a database can also be useful for other purposes. For example, it could be used to determine the effect on the results of short-term tests of different prevalences of carcinogens. This, in turn, will influence the testing strategy used to detect carcinogens. Such an analysis shows (Table 13) that there is a considerable proportion of false positive results to be expected, as evidenced by the fact that when the prevalence of carcinogens is 0%, we can expect 14 and 47% of the chemicals to respond positively in the Salmonella mutagenicity and sister chromatid assays, respectively (Table 13). Moreover, an unacceptably high rate of false negatives is to be expected as well, for a population of only carcinogens (100% prevalence) is expected to yield only a 44% rate of positive responses in the Salmonella mutagenicity assay.

# Conclusion

Powerful computer-based expert systems to study SAR are now available. However, as illustrated here, these methodologies are greatly dependent on the nature and organization of databases. Moreover, it has been demonstrated that databases need not be extensive for SAR analysis. Thus the present studies indicate that predictive toxicology is possible.

This investigation was supported by the National Institute of Environmental Health Sciences (ES04659) and the U.S. Environmental Protection Agency (R815488).

## REFERENCES

1. Klopman, G., and Rosenkranz, H. S. Structural requirements for the mutagenicity of environmental nitroarenes. Mutat. Res. 126: 227–238 (1984).
2. Klopman, G., Frierson, M. R., and Rosenkranz, H. S. The structural basis of the mutagenicity of chemicals in *Salmonella typhimurium*: the Gene-Tox Data Base. Mutat. Res. 228: 1–50 (1990).
3. Rosenkranz, H. S., Ennever, F. K., Dimayuga, M., and Klopman, G. Significant differences in the structural basis of the induction of sister chromatid exchanges and chromosomal aberrations in Chinese hamster ovary cells. Environ. Mol. Mutagen. 16: 149–177 (1990).
4. Klopman, G., Kalos, A. N., and Rosenkranz, H. S. An artificial intelligence study of the structure–activity relationships of non-fused ring nitroarenes and related compounds. Mol. Toxicol. 1: 61–81 (1987).
5. Klopman, G., Frierson, M. R., and Rosenkranz, H. S. Computer analysis of toxicological databases: mutagenicity of aromatic amines in Salmonella tester strains. Environ. Mutagen. 7: 625–644 (1985).
6. Rosenkranz, H. S., McCoy, E. C., Frierson, M., and Klopman, G. The role of DNA sequence and structure of the electrophile on the mutagenicity of nitroarenes and arylamine derivatives. Environ. Mutagen. 7: 645–653 (1985).
7. Rosenkranz, H. S., and Klopman, G. The structural basis of the mutagenicity of chemicals in *Salmonella typhimurium*: the National Toxicology Program data base. Mutat. Res. 228: 51–80 (1990).
8. Tennant, R. W., Margolin, B. H., Shelby, M. D., Zeiger, E., Haseman, J. K., Spalding, J., Caspary, W., Resnick, M., Stasiewicz, S., Anderson, B., and Minor, R. Prediction of chemical carcinogenicity in rodents from in vitro genotoxicity assays. Science 236: 933–941 (1987).
9. Ashby, J., and Tennant, R. W. Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. Mutat. Res. 204: 17–115 (1988).
10. Ashby, J., Tennant, R. W., Zeiger, E., and Stasiewicz, S. Classification according to chemical structure, mutagenicity to Salmonella and level of carcinogenicity of a further 42 chemicals tested for carcinogenicity by the U.S. National Toxicology Program. Mutat. Res. 223: 73–103 (1989).
11. Balls, M., Blaauboer, B., Brusick, D., Frazier, J., Lamb, D., Pemberton, M., Reinhardt, C., Roberfroid, M., Rosenkranz, H., Schmid, B., Speilmann, H., Stammati, A.-L., and Walum, E. Report and recommendations of the CAAT/ERGATT Workshop on the Validation of Toxicity Test Procedures. Alt. Lab. Anim. 18: 313–337 (1990).
12. Ray, V. A., Kier, L. D., Kannan, K. L., Haas, R. T., Auletta, A. E., Wassom, J. S., Nesnow, S., and Waters, M. D. An approach to identifying specialized batteries of bioassays for specific classes chemicals: class analysis using mutagenicity and carcinogenicity relationships and phylogenetic concordance and disconcordance patterns. 1. Composition and analysis of the overall database. A report of Phase II of the U.S. Environmental Protection Agency Gene-Tox Carcinogen Data Base. Mutat. Res. 185: 197–241 (1987).
13. Nesnow, S., Argus, M., Bergman, H., Chu, K., Frith, C., Helmes, T., McGaughey, R., Ray, V., Slaga, T. J., Tennant, R., and Weisburger, E. Chemical carcinogens: a review and analysis of the literature of selected chemicals and the establishment of the Gene-Tox Carcinogen Data Base. Mutat. Res. 185: 1–195 (1987).
14. Ashby, J. A. Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. Environ. Mutagen. 7: 919–921 (1985).
15. Rosenkranz, H. S., and Klopman, G. Structural basis of carcinogenicity in rodents of genotoxicants and non-genotoxicants. Mutat. Res. 228: 105–124 (1990).
16. Rosenkranz, H. S., and Klopman, G. Natural pesticides present in edible plants are predicted to be carcinogenic. Carcinogenesis 11: 349–353 (1990).
17. Klopman, G., and Rosenkranz, H. S. Computational alternative to the experimental determination of mutagenicity in Salmonella. Submitted.
18. National Academy of Sciences. Toxicity Testing. Stategies to Determine Needs and Priorities. National Academy Press, Washington, DC, 1984.
19. Rosenkranz, H. S., and Klopman, G. Structure activity based predictive toxicology: an efficient and economical methods for generating non-congeneric data bases. Mutagenesis, in press.
20. Rosenkranz, H. S., and Klopman, G. Mechanistic insights gained from an analysis of carcinogenic polycyclic aromatic hydrocarbons with the Computer Automated Structure Evaluation system. J. Am. Coll. Toxicol. 8: 1091–1101 (1989).
21. Gold, L. S., Sawyer, C. B., R., Magaw, R., Backman, G. M., de Veciana, M., Levinson, R., Hooper, N. K., Havender, W. R., Bernstein, L., Peto, R., Pike, M. C., and Ames, B. N. A Carcinogenic Potency Database of the standardized results of animal bioassays. Environ. Health Perspect. 58: 9–319 (1984).
22. Gold, L. S., de Veciana, M., Backman, G. M., Magaw, R., Lopipero, P., Smith, M., Blumenthal, M., Levinson, R., Bernstein, L., and Ames, B. N.

Chronological supplement to the Carcinogenic Potency Database: standardized results of animal bioassays published through December 1982. Environ. Health Perspect. 67: 161–200 (1986).

23. Gold, L. S., Slone, T. H., Backman, G. M., Magaw, R., Da Costa, M., Lopipero, P., Blumenthal, M., and Ames, B. N. Second chronological supplement to the Carcinogenic Potency Database: standardized results of animal bioassays published through December 1984 and by the National Toxicology Program through May 1986. Environ. Health Perspect. 74: 237–329 (1987).
24. Peto, R., Pike, M. C., Berstein, L., Gold, L. S., and Ames, B. N. The $TD_{50}$: a proposed general convention for the numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments. Environ. Health Perspect. 58: 1–8 (1984).
25. Rosenkranz, H. S., Ennever, F. K., and Klopman, G. Relationship between carcinogenicity in rodents and the induction of sister chromatid exchanges and chromosomal aberrations in Chinese hamster ovary cells. Mutagenesis 5: 559–571 (1990).